A Novel Nonparametric Density Estimator

Z. I. Botev*
The University of Queensland
Australia

Abstract

We present a novel nonparametric density estimator and a new data-driven bandwidth selection method with excellent properties. The approach is inspired by the principles of the generalized cross entropy method. The proposed density estimation procedure has numerous advantages over the traditional kernel density estimator methods. Firstly, for the first time in the nonparametric literature, the proposed estimator allows for a genuine incorporation of prior information in the density estimation procedure. Secondly, the approach provides the first data-driven bandwidth selection method that is guaranteed to provide a unique bandwidth for any data. Lastly, simulation examples suggest the proposed approach outperforms the current state of the art in nonparametric density estimation in terms of accuracy and reliability.

Keywords kernel density estimation, bandwidth selection, partial differential equation, heat kernel, Cross Entropy method

*Supported by the Australian Research Council, under grant number DP0558957.

1. Background. Suppose we are given continuous data $\mathcal{X}_N \equiv \{X_1, \ldots, X_N\}$ on $\mathscr{X} \subseteq \mathbb{R}$. Assume that the data are i.i.d realizations from an unknown continuous pdf f, i.e., $X_1, \ldots, X_N \overset{i.i.d}{\sim} f$. The problem is to estimate f from the data $\mathcal{X}_N$, using as few assumptions as possible. The empirical pdf $\Delta(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_i)$ is not a good model for the continuous f, because Δ is not a continuous function; see [7]. The standard approach for estimating f is to use the *Kernel Density Estimator* (KDE):

$$g(x; t \,|\, \mathcal{X}_N) = \mathbb{E}_\Delta[K(x, X; t)] = \frac{1}{N}\sum_{i=1}^{N} K(x, X_i; t), \tag{1}$$

where K is a *kernel* function. The most common choice for a kernel is the Gaussian pdf with mean θ and variance t: $K(x, \theta; t) = \frac{1}{\sqrt{2\pi t}}\, \mathrm{e}^{-\frac{(x-\theta)^2}{2t}}$. The only unknow in (1) is the parameter t, which we call the *bandwidth*. A lot of research has focused on the optimal choice of the bandwidth, because the performance of g as an estimator of f depends crucially on its value; see, e.g., [7] and the references therein. The classical measure of performance of the estimator (1) is the *Mean Integrated Squared Error* (MISE):

$$\mathrm{MISE}\{g\}(t) = \mathbb{E}_f\left[\int [g(x; t \,|\, \mathcal{X}_N) - f(x)]^2 dx\right],$$

where g is viewed as a function of the random sample $\mathcal{X}_N$, and the expectation operator applies to $\mathcal{X}_N$. To simplify the notation, we omit from now on any explicit reference to the dependence of g on $\mathcal{X}_N$. Let $\mathrm{Bias}\{g\}(x; t) = \mathbb{E}_f[g(x; t)] - f(x)$ denote the point-wise bias of (1). Then the MISE can be written as:

$$\mathrm{MISE}\{g\}(t) = \int \mathrm{Bias}^2\{g\}(x; t)\, dx + \int \mathrm{Var}_f[g(x; t)]\, dx.$$

Unfortunately, MISE depends on the unknown f in a complicated way. This is why, instead of considering the MISE directly, one considers the asymptotic approximation of MISE as the sample size $N \to \infty$. It has been shown [6],[7] that under the assumptions that t depends on N such that $\lim_{N\to\infty} t = 0$, $\lim_{N\to\infty} N\sqrt{t} = \infty$ and f'' is a continuous square integrable function and K is a Gaussian kernel, the estimator (1) is consistent and has integrated squared bias and integrated variance given by:

$$||\mathrm{Bias}\{g\}(\cdot; t)||^2 = \frac{1}{4}t^2||f''||^2 + o(t^2),\ N \to \infty \tag{2}$$

$$\int \mathrm{Var}_f[g(x; t)]\, dx = \frac{1}{2N\sqrt{\pi t}} + o((N\sqrt{t})^{-1}),\ N \to \infty. \tag{3}$$

Here $||\cdot||$ denotes the standard L^2 norm on $\mathscr{X}$. It follows that the first-order asymptotic approximation of MISE, denoted AMISE, is given by

$$\mathrm{AMISE}\{g\}(t) = \frac{1}{4}t^2||f''||^2 + (2N\sqrt{\pi t})^{-1}. \tag{4}$$

The optimal bandwidth is defined as the minimizer of (4): ${}_*t = \left(2N\sqrt{\pi}\,||f''||^2\right)^{-2/5}$. Unfortunately AMISE and ${}_*t$ still depend on the unknown f through the functional $||f''||^2$, rendering the whole procedure rather dubious. The simplest and naive solution is to *plug-in* a parametric estimate of f'' in $||f''||^2$, where f is usually assumed to be a Gaussian pdf with mean and variance estimated from the data. This assumption can be very wrong rendering the bandwidth estimate useless. More high-tech approaches such as the Sheather-Jones (SJ) estimation method [6] estimate $||f''||^2$ via $||g''(\cdot; t_s)||^2$. Here $g(\cdot; t_s)$ is a *second stage* KDE with a different bandwidth t_s, which Sheather and Jones show to be asymptotically related to ${}_*t$ itself. We

denote the KDE of the SJ method $g_{\text{SJ}}(\cdot; t_{\text{SJ}})$, where t_{SJ} is the optimal SJ bandwidth [6] and the kernel is Gaussian. Regrettably the dependence of the second stage estimator $g(\cdot; t_s)$ on unknown functionals of f does not go away. Thus the SJ procedure again eventually assumes a Gaussian model for $\mathcal{X}_N$. The accomplishment of the SJ method is that the effect of the normality assumption on the reliability of the bandwidth selection procedure is significantly diminished, making the approach widely regarded as the current state of the art. Nevertheless, simple examples can be constructed [2] in which the target f is a bimodal density and the SJ method fails because the assumption that $\mathcal{X}_N$ is normally distributed is blatantly wrong. Such an example will be provided in the last part of this paper.

2. A Novel Estimator. We now present the new *Generalized Cross Entropy* (GCE) estimator of the form (1). The GCE estimator will motivate a new bandwidth selection method which obviates the above mentioned problems. For a motivation of the estimator and its relation to Entropy see [1]. For simplicity consider $\mathscr{X} \equiv [0, 1]$. There is no loss of generality since $\mathcal{X}_N$ can always be mapped onto the interval $[0, 1]$ by an invertible transformation. Let $p(x) > 0,\ x \in \mathscr{X}$ represent the *prior* density of $\mathcal{X}_N$. The prior density represents all the a priori information about the distribution of $\mathcal{X}_N$. If we have no prior information about the distribution of $\mathcal{X}_N$, we set $p \equiv 1$. For the rest of the paper let K be the *heat kernel* on $\mathscr{X}$ with *spectral representation*:

$$K(x, \theta; t) = p(x) \sum_{k=0}^{\infty} \mathrm{e}^{\lambda_k t/2} \phi_k(x) \phi_k(\theta), \tag{5}$$

where $\{1, \phi_1, \phi_2, \ldots\}$ and $\{0 = \lambda_0 > \lambda_1 > \lambda_2 > \ldots\}$ are the (normalized) eigenfunctions and eigen-values of the *regular Sturm-Liouville* boundary value problem on $[0, 1]$:

$$\phi_k''(x) = \lambda_k\, p(x) \phi_k(x), \qquad \phi_k'(0) = \phi_k'(1) = 0, \quad k = 0, 1, \ldots. \tag{6}$$

It is well known [4] that $\{\phi_k\}$ forms a complete orthonormal basis (with respect to weight p) for $L^2(0, 1)$. Our proposed estimator is (1) with K given by (5), i.e.:

$$g_{\text{GCE}}(x; t) = p(x) \sum_{k=0}^{\infty} \mathrm{e}^{\lambda_k t/2} \hat{\varphi}_k\, \phi_k(x), \text{ with } \hat{\varphi}_k = \mathbb{E}_\Delta[\phi_k(X)] = \frac{1}{N} \sum_{i=1}^{N} \phi_k(X_i). \tag{7}$$

It can be shown, using arguments similar to those given in [7], that the asymptotic behavior of (7) is given by:

Theorem 1 Under the conditions that f'' is continuous and square integrable with $f'(0) = f'(1) = 0$ and $\lim_{N\to\infty} t = 0$, $\lim_{N\to\infty} \sqrt{t} N = \infty$, we have:

$$\begin{aligned} ||\text{Bias}\{g_{\text{GCE}}\}(\cdot; t)||^2 &= \frac{1}{4} t^2 ||(f/p)''||^2 + o(t^2), \quad N \to \infty, & (8)\\ \int \text{Var}_f[g_{\text{GCE}}(x; t)]\, dx &= \frac{1}{2N\sqrt{\pi t}} + o((N\sqrt{t})^{-1}), \quad N \to \infty, & (9)\\ \text{AMISE}\{g_{\text{GCE}}\}(t) &= \frac{1}{4} t^2 ||(f/p)''||^2 + (2N\sqrt{\pi t})^{-1}. & (10) \end{aligned}$$

Therefore

$$\left(2N\sqrt{\pi}\, ||(f/p)''||^2\right)^{-2/5} = \operatorname*{argmin}_{t>0} \text{AMISE}\{g_{\text{GCE}}\}(t) \tag{11}$$

is the AMISE optimal bandwidth of g_{GCE}. Note that the condition $f'(0) = f'(1) = 0$ is not restrictive since $\mathcal{X}_N$ can always be mapped to $[0, 1]$ in such a way that the

first and last order statistic are at some distance away from the boundary rendering any boundary effects asymptotically negligible. The asymptotic integrated variance (3) of the traditional KDE and the asymptotic integrated variance (9) of g_{GCE} are the same, but note that the asymptotic norm of the bias (8) of g_{GCE} is different from (2) due to the introduction of the prior p. It is now clear that for a given fixed t, $\text{AMISE}\{g_{\text{GCE}}\}(t) \leq \text{AMISE}\{g_{\text{SJ}}\}(t)$ if and only if $||\text{Bias}\{g_{\text{GCE}}\}(\cdot;t)|| \leq ||\text{Bias}\{g_{\text{SJ}}\}(\cdot;t)||$, which in turn is equivalent to $||\big(f/p\big)''|| \leq ||f''||$, i.e., the prior p must be close to f. If, for example, $p \equiv f$, the asymptotic norm of the bias of g_{GCE} is zero. And if $p \equiv 1$, corresponding to no prior information, then it is identical to the norm of the bias (2) of the traditional KDE. Thus one advantage of the proposed estimator is that it allows us to genuinely incorporate prior information. This prior information can potentially improve the AMISE performance of the estimator. Another advantage of the method is that, even without the availability of prior information, the estimator motivates a new data-driven bandwidth selection method as explained in the next part.

3. A New Bandwidth Selection Procedure. Our reliable bandwidth selection rule is motivated by the following arguments. First, straightforward integration of (7) yields the following estimator of the cdf F of f:

$$G(x;t \,|\, \mathcal{X}_N) = x + \sum_{k=1}^{\infty} \frac{\mathrm{e}^{\lambda_k t/2}\hat{\varphi}_k}{\lambda_k}\, \phi_k'(x). \tag{12}$$

$\{\phi_k'\}$ is an orthogonal set [4] with respect to weight 1 and $||\phi_k'||^2 = -\lambda_k$. Next, it is easy to show [3] that under the conditions of theorem 1, the asymptotic behavior as $N \to \infty$ of the MISE of the estimator (12) is:

$$\mathbb{E}_f\bigg[\int_0^1 [G(x;t \,|\, \mathcal{X}_N) - F(x)]^2 dx\bigg] = \frac{1}{N}\int_0^1 F(x)(1-F(x))dx - \frac{\sqrt{t/\pi}}{N} + o(\sqrt{t}). \tag{13}$$

Lastly, substituting G in (13) for (12) and ignoring the asymptotically negligible terms of order $o(\sqrt{t})$, one obtains, after simplification:

$$\sum_{k=1}^{\infty} \frac{1}{-\lambda_k}\mathbb{E}_f\bigg[\Big(\mathrm{e}^{\lambda_k t/2}\hat{\varphi}_k - \mathbb{E}_f[\phi_k(X)]\Big)^2\bigg] + \frac{\sqrt{t/\pi}}{N} = \frac{1}{N}\int_0^1 F(x)(1-F(x))dx. \tag{14}$$

At first sight it seems that the asymptotic MISE approximation (14) suffers from the same problems as the asymptotic MISE (4) of g in that it depends on the unknown F and f. The important difference however is that, unlike the $\text{AMISE}\{g\}(t)$, (14) allows us to substitute both f and F with their unique unbiased estimators, namely, the empirical pdf Δ and cdf $\hat{F}$ respectively. Thus the substitution $f \to \Delta$ and $F \to \hat{F}$ yields, after simplification, the *stochastic counterpart* of (14):

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}_\Delta[\hat{\varphi}_k^2]}{-\lambda_k}\Big(\mathrm{e}^{\lambda_k t/2} - 1\Big)^2 + \frac{\sqrt{t/\pi}}{N} = \frac{1}{N}\int_0^1 \hat{F}(x)(1-\hat{F}(x))dx. \tag{15}$$

Recalling that $\lambda_k < 0$, $\forall k \geq 1$, it follows that the left-hand side of (15) is a monotonic function of t. This immediately yields:

Theorem 2 For a finite N and any empirical distribution function $\hat{F}$, there exists a unique root t^* of the non-linear equation (15), such that $0 < t^* < \infty$.

We define the unique t^* as our novel and reliable optimal bandwidth estimate. Theorem 2 is one argument in our contention that the proposed method has practical

performance second to none in the existing literature. Theorem 2 appears to be the first uniqueness result for a hi-tech bandwidth selection method.

Theoretically the root of (15), t^*, provides an estimate of the asymptotically optimal bandwidth for the estimator (12) but not for (7) directly. A standard solution is to use t^* as the bandwidth for the second stage KDE of the functional $||(f/p)''||^2$ and then employ the AMISE optimal bandwidth (11). In other words, we estimate $||(f/p)''||^2$ via $||(g_{\text{GCE}}(\cdot;t^*)/p)''||^2$ and then use

$$t_{\text{GCE}} = \left(2N\sqrt{\pi}\,||\,(g_{\text{GCE}}(\cdot;t^*)/p)''\,||^2\right)^{-2/5}$$

as a plug-in estimate of the AMISE optimal bandwidth (11). Thus given any data $\mathcal{X}_N$ and any prior information about $\mathcal{X}_N$ in the form of a prior density p, the GCE model for $\mathcal{X}_N$ is the estimator (7) with $t_{\text{GCE}} = \left(2N\sqrt{\pi}\,||\,(g_{\text{GCE}}(\cdot;t^*)/p)''\,||^2\right)^{-2/5}$ and t^* being the root of (15). In many practical situations we do not have any prior information (i.e., $p \equiv 1$). We can, however, always take advantage of the possibility of incorporating prior information in (7) by first performing a preliminary (pilot) estimation step of f and then using the pilot estimate as a prior in estimator (7). This idea is summarized in the following density estimation procedure:

Algorithm 1 (GCE estimator with pilot density estimation)

1. Given the data $\mathcal{X}_N$ with no prior information, let (7) with $p \equiv 1$ and bandwidth $t_{\text{GCE}} = \left(2N\sqrt{\pi}\,||g''(\cdot;t^*)||^2\right)^{-2/5}$ be the pilot model for $\mathcal{X}_N$. Here t^* is the root of (15).

2. Using the same t_{GCE}, compute the final estimate (7) with the prior p identically equal to the pilot estimate from step 1.

Note that the algorithm is heuristic in the sense that we do not have any genuine prior information but instead utilize the same $\mathcal{X}_N$ to construct p. This is the reason why we use the same $t_{\text{GCE}} = \left(2N\sqrt{\pi}\,||g''_{\text{GCE}}(\cdot;t^*)||^2\right)^{-2/5}$, corresponding to $p \equiv 1$, in both step 1 and step 2. The numerical aspect of the algorithm does not present any problems. Since the estimator (7) satisfies the PDE $\frac{\partial}{\partial t}g(x;t) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\frac{g(x;t)}{p(x)}\right)$ with boundary condition $\frac{\partial}{\partial x}\left(\frac{g(x;t)}{p(x)}\right)\Big|_{x=0} = \frac{\partial}{\partial x}\left(\frac{g(x;t)}{p(x)}\right)\Big|_{x=1} = 0$ and initial condition $g(x;0) = \Delta(x)$, we can use any of the standard numerical methods for PDEs to construct the GCE estimator (7) and the solution of (15). Our implementation uses the spectral Galerkin method [4]. In the PDE context we can interpret t_{GCE} as the optimal stopping time of an evolution process governed by a diffusion PDE and the kernel (5) as the *Green's function* of the PDE [4]. The kernel (5) has the property that for a given t, it applies more smoothing in regions of small p (low density) and less smoothing in regions of large p (high density).

4. Simulation Experiments. We compared the GCE estimator of Algorithm 1 against the popular SJ estimator, [6], using Ratio $= ||g_{\text{GCE}} - f||^2/||g_{\text{SJ}} - f||^2$ as our criterion, i.e., the ratio of the (exact) integrated squared error of the GCE estimator to the integrated squared error of the SJ estimator. We used Professor Steve Marron's Matlab implementation of the SJ method freely available at: http://www.stat.unc.edu/faculty/marron/marron_software.html . Table 1 shows the average results over 10 independent trials for seven different test cases. The second column of table 1 displays the target density and the third column shows the sample size used for the experiments. The standard Gaussian mixture test problems are taken from [5]. Figure 1 shows the results of a single simulation for the test problems 1 through 4 in Table 1. Note how in test case 2 the GCE estimator has fewer spurious modes in regions of low density and a suitably peaked mode in the high density

Test Problem	target density $f(x)$	N	Ratio	winner
1	$\frac{1}{2}\mathsf{N}\left(0,\left(\frac{1}{10}\right)^2\right)+\frac{1}{2}\mathsf{N}(5,1)$	200	0.25	GCE
2	$\frac{2}{3}\mathsf{N}(0,1)+\frac{1}{3}\mathsf{N}\left(0,\left(\frac{1}{10}\right)^2\right)$	300	0.60	GCE
3	$\frac{1}{2}\mathsf{N}(0,1)+\sum_{k=0}^{4}\frac{1}{10}\mathsf{N}\left(\frac{k}{2}-1,\left(\frac{1}{10}\right)^2\right)$	400	0.76	GCE
4	$\sum_{k=0}^{7}\frac{1}{8}\mathsf{N}\left(3\left(\left(\frac{2}{3}\right)^k-1\right),\left(\frac{2}{3}\right)^{2k}\right)$	400	0.58	GCE
5	$\frac{3}{4}\mathsf{N}(0,1)+\frac{1}{4}\mathsf{N}\left(\frac{3}{2},\left(\frac{1}{3}\right)^2\right)$	400	0.87	GCE
6	Log-Normal with $\mu=0$ and $\sigma=1$	400	0.71	GCE
7	Exponential with mean $\mu=1$	400	0.39	GCE

Table 1: Results over 10 independent simulation experiments.

region. The improvement in the GCE estimator is due to the adaptive application of different amounts of smoothing in regions of different density. Test problem 1 (separated bimodal density) demonstrates the fact that the SJ estimator does not pass the bimodality test [2] whilst the GCE estimator does. The bi-modality test [2] consists of testing the performance of a given estimation procedure on a bimodal target density f with the two modes at some distance from each other. It has been demonstrated [2] that by separating the modes of the target density enough, the SJ method can be made to perform arbitrarily poorly because the eventual assumption of normality within the SJ procedure fails. No such problems exist for the proposed GCE estimator. In conclusion we point out that the GCE approach yields smaller integrated squared error than the SJ approach in all of the test cases. As future research one could explore the possibility of utilizing the optimal t^* within the SJ method itself. The optimal t^* could be used to dispense with the currently used normality assumption. The result could be the removal of the bimodality weakness of the SJ method and overall improvement in reliability.

References

[1] Z. I. Botev. *Stochastic Methods for Optimization and Machine Learning*. ePrintsUQ, http://eprint.uq.edu.au/archive/00003377/, Technical Report, 2005.

[2] Luc Devroye. Universal smoothing factor selection in density estimation: theory and practice. *Sociedad de stadistica e Investigacian Operativa*, 6(2), 1997.

[3] M. C. Jones. The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*, 9:129–132, 1990.

[4] Stig Larsson and Vidar Thomee. *Partial Differential Equations with Numerical Methods*. Springer, 2003.

[5] J. S. Marron and M. P. Wand. Exact mean integrated error. *The Annals of Statistics*, 20:712–736, 1992.

[6] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J.R.Statist.Soc.B*, 53:683–690, 1991.

[7] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

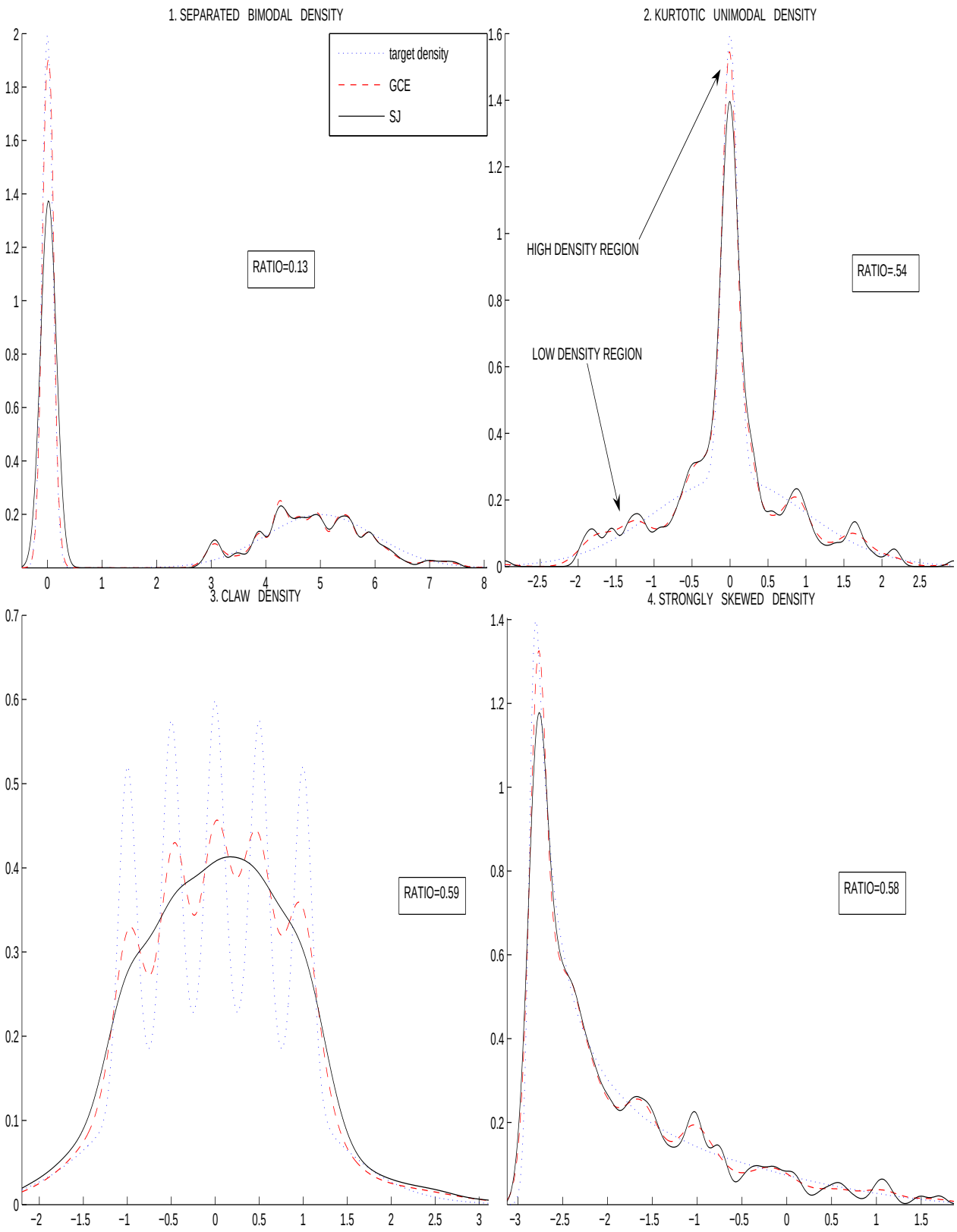

Figure 1: Simulation results for test problems 1 through 4 in Table 1.